
Slik kan man studere bruken av medisinske termer

SPRÅKSPALTEN

LARS G. BAGØIEN JOHNSEN

lars.johnsen@nb.no

Lars G. Bagøien Johnsen har doktorgrad i lingvistikk og er forskningsbibliotekar ved Nasjonalbiblioteket i Oslo.

Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Nasjonalbibliotekets digitaliserte samling av tekster er tilgjengelig for alle som vil studere medisinsk terminologi.



Tilfeldige bokomslag fra Nasjonalbiblioteket sin samling. Skjermdumper fra nb.no. S sammensatt av Tidsskriftet.

Data om medisinsk språkbruk gjør det mulig å si noe om hvordan ord brukes, hvordan alternativer forholder seg til hverandre og eventuelle endringer i bruk (1). Digitalt tilgjengelige tekster fra Nasjonalbibliotekets digitale bibliotek kan brukes som kilder til slike studier.

Ved Nasjonalbiblioteket er hele den tekstbaserte norske kulturarven digitalisert og gjort tilgjengelig både for lesing og for dataanalyse (2). Over 600 000 bøker, over fire millioner aviser og ca. 100 000 tidsskrifter er tilgjengelig digitalt. For bøker er samlingen nesten komplett, og alle aviser og tidsskrifter er under kontinuerlig digitalisering. Tilgjengelige data om tekstene omfatter blant annet frekvenser og statistikk basert på kobling mellom metadata og ord.

Den digitale infrastrukturen består også av algoritmer som kan analysere tekster og gjøre søk i dem og på annet vis trekke ut informasjon. I tillegg til nettbiblioteket har Nasjonalbiblioteket et laboratorium som tilbyr programmatisk innganger til data via enkelte programmeringsspråk og apper som kan tilpasses forskningsprosjekter (2).

Korpus

Med *korpus* forstår vi en samling tekster som har et sett egenskaper som gjør at vi kan trekke slutninger om bruksmønstre, betydning og grammatikk, basert på hvordan ordene opptrer i det (2). I leksikografisk og terminologisk arbeid er vi særlig ute etter ords egenskaper, ord som eksisterer i flere fagfelt og ord som har forskjellige egenskaper. Men ofte er enkelte termer typiske for et bestemt fagfelt, og da er ikke alltid korpusets definisjon relevant. Men medisinen benytter termer som også har en allmenn bruk, for eksempel *øye* og *hode*, og i slike tilfeller er det viktig å ha et korpus som definerer bruken av ordene i en medisinsk kontekst.

Alle bøker og tidsskrifter er utstyrt med to typer metadata. Det ene er kataloginformasjon, slik som publikasjonsår, tittel etc. Det andre er klassifikasjon av innholdet, som gjøres med emneord og et desimalsystem, som for medisin er desimaltall 610*. Vi kan derfor bygge et korpus med utgangspunkt i dette desimalsystemet.

En tradisjonell korpuslingvistisk metode er å først definere korpuset med metadata og så studere fordelingen av ord inne i korpuset. Ved å danne et korpus med utgangspunkt i medisinsk litteratur vil vi kunne påberope oss en viss autoritet i påstander om bruken av visse ord, både fra bøker og/eller medisinske tidsskrifter. Den grunnleggende infrastrukturen består av databaser som ligger på en server som er forskjellig fra den maskinen man arbeider fra. Det er bygget et programbibliotek som gjør det mulig å lage brukervennlige grensesnitt for leksikografer.

Konkordans og kollokasjon

En slik undersøkelse gjøres gjerne i to deler. Først ser man på *konkordans*, dvs. bruksfrekvens for et ord, deretter *kollokasjon*, som kan defineres som en forbindelse av to eller flere ord som vanligvis opptrer sammen (3), for eksempel *gjøre fremskritt* og *felle en dom*. Dette kan også kalles aggregerte konkordanser. Man kan også bruke begrepet i en litt videre betydning enn dette, slik at kollokatene for et ord ikke trenger å stå ved siden av ordet, men at det holder at de befinner seg innenfor en kontekst og er assosiert.

Med kollokasjoner kan man se på større eller mindre kontekstvinduer og også begrense dem til bare noen få ord til høyre eller venstre for å undersøke sammenhengen mellom ord og hva de typisk opptrer med. Kollokatene, altså de ordene som antas å være knyttet til et *målord*, og eventuelle endringer i dem, kan fortelle oss noe om forskjellige betydninger. Kollokatene kan gis en

relevansverdi basert på frekvensforskjellen mellom forekomster i konkordansene (kontekst) og korpuset. Om ordet har en frekvens på 2 % av alle ordene i kontekst for målordet, samtidig som det opptrer i 0,5 % i korpuset ellers, kan vi si at det opptrer fire ganger så ofte med målordet enn utenfor. Frekvensforskjellen gir et mål på assosiasjonen mellom kollokatet og målordet: Jo høyere, jo sterkere assosiert.

Ressursene som tilbys ved Nasjonalbiblioteket, kan benyttes i medisinskterminologisk arbeid. Informasjonsflyten kan gå fra ord til tekst og så til klassifikasjonsdata, og omvendt fra klassifikasjon til tekster og så til ord og termer. Selv om ikke alt som er publisert innen medisin er gjort tilgjengelig, er det likevel store mengder informasjon som kan hentes ut.

*Denne artikkelen er en forkortet utgave av et kapittel (4) i boken *Pasientjournalen – språk, dokumentasjon og helsekompetanse* (5).*

REFERENCES

1. Birkenes MB, Johnsen LG, Lindstad AM et al. From digital library to n-grams: NB N-gram. <https://ep.liu.se/ecp/109/039/ecp15109039.pdf> Lest 20.11.2023.
2. Nasjonalbiblioteket. DH-LAB. <https://www.nb.no/dh-lab/> Lest 20.11.2023.
3. Sinclair J. red. Corpus, concordance, collocation. Oxford: Oxford University Press, 1991.
4. Johnsen LGB. Å bruke språkteknologi til å undersøke medisinske ord. Michael 2023; 20 (Suppl 31): 133–42.
5. Nylenna M, Hem E, Husom N. red. Pasientjournalen – språk, dokumentasjon og helsekompetanse. Michael 2023; 20 (Suppl 31): 7–207.

Publisert: 7. mai 2024. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.23.0837
Opphavsrett: © Tidsskriftet 2025 Lastet ned fra tidsskriftet.no 27. desember 2025.